Estimating Recombination Breakpoint Locations Using Hidden Markov Models and Reversible Jump MCMC

Alex Webb[1], Chris Holmes[1]

[1] Department of Statistics, University of Oxford, UK

Recombination is one of the fundamental processes for creating diversity and yet traditional phylogenetic inference assumes that the history of a set of taxa can be explained by a single tree. Since this assumption will often be violated the phylogenies inferred may be incorrect and at worst will be meaningless trees that try to 'average' phylogenies for different regions of the DNA sequence. Accounting for recombination will allow improved phylogenetic inference and, in the case of pathogens such as HIV, this may help lead to more targeted drug development.
We have developed a Hidden Markov Model to allow for different phylogenies to be inferred for different parts of our DNA alignment. The states of our HMM represent unrooted phylogenies and each site in the alignment is a realization of a mutational process along the branches of one of the phylogenies. Using the structure of HMMs we can efficiently integrate over all topologies, and all mutational configurations given those topologies. The topologies used in the model are updated using a reversible jump Markov chain Monte Carlo approach. This allows us to introduce new states to the model as they are required - this allows the data to dictate how many phylogenies are required to explain the data.
Testing the method on simulated data has produced encouraging results and we have applied it to HIV data and to SNP data from 15 inbred strains of laboratory mice. However, our method could be applied to a wide variety of problems in order to explore the possibility of recombination in the history of the dataset.